

(19)



Europäisches Patentamt

European Patent Office

Office européen des brevets



(11)

**EP 0 508 604 B1**

(12)

## EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention  
of the grant of the patent:  
**17.11.1999 Bulletin 1999/46**

(51) Int Cl.<sup>6</sup>: **G06F 3/06**, G06F 13/42,  
G06F 11/10

(21) Application number: **92302098.6**

(22) Date of filing: **11.03.1992**

### (54) **Disk array controller for data storage system**

Speicherplattenanordnungsteuerungsvorrichtung für eine Datenspeicherungsanordnung

Dispositif de commande d'un réseau de disques pour un système de stockage de données

(84) Designated Contracting States:  
**DE FR GB**

(30) Priority: **13.03.1991 US 668660**

(43) Date of publication of application:  
**14.10.1992 Bulletin 1992/42**

(73) Proprietors:  
• **NCR International, Inc.**  
**Dayton, Ohio 45479 (US)**  
• **HYUNDAI ELECTRONICS AMERICA**  
**Milpitas, California 95035 (US)**  
• **Symbios, Inc.**  
**Fort Collins, Colorado 80525 (US)**

(72) Inventors:  
• **Jibbe, Mahmoud K.**  
**Wichita, KS 67208 (US)**

• **McCombs, Craig C.**  
**Wichita, KS 67213 (US)**  
• **Thompson, Kenneth J.**  
**Wichita, KS 67226 (US)**

(74) Representative: **Gill, David Alan et al**  
**W.P. Thompson & Co.,**  
**Celcon House,**  
**289-293 High Holborn**  
**London WC1V 7HU (GB)**

(56) References cited:  
**EP-A- 0 294 287** **EP-A- 0 320 107**  
**EP-A- 0 369 707** **US-A- 4 899 342**  
**US-A- 4 914 656**

**EP 0 508 604 B1**

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

## Description

[0001] This invention relates to data storage systems of the kind including host bus means adapted to be connected to a host device, and disk drive means adapted to be connected to a plurality of disk drive devices.

[0002] Recent and continuing increases in computer processing power and speed, in the speed and capacity of primary memory, and in the size and complexity of computer software has resulted in the need for faster operating, larger capacity secondary memory storage devices; magnetic disks forming the most common external or secondary memory storage means utilized in present day computer systems. Unfortunately, the rate of improvement in the performance of large magnetic disks has not kept pace with processor and main memory performance improvements. However, significant secondary memory storage performance and cost improvements may be obtained by the replacement of single large expensive disk drives with a multiplicity of small, inexpensive disk drives interconnected in a parallel array, which to the host appears as a single large fast disk.

[0003] Several disk array design alternatives were presented in an article titled "A Case for Redundant Arrays of Inexpensive Disks (RAID)" by David A. Patterson, Garth Gibson and Randy H. Katz; University of California Report No. UCB/CSD 87/391, December 1987. This article discusses disk arrays and the improvements in performance, reliability, power consumption and scalability that disk arrays provide in comparison to single large magnetic disks.

[0004] Five disk array arrangements, referred to as RAID levels, are described in the article. The first level RAID comprises N disks for storing data and N additional "mirror" disks for storing copies of the information written to the data disks. RAID level 1 write functions require that data be written to two disks, the second "mirror" disk receiving the same information provided to the first disk. When data is read, it can be read from either disk.

[0005] RAID level 3 systems comprise one or more groups of N+1 disks. Within each group, N disks are used to store data, and the additional disk is utilized to store parity information. During RAID level 3 write functions, each block of data is divided into N portions for storage among the N data disks. The corresponding parity information is written to a dedicated parity disk. When data is read, all N data disks must be accessed. The parity disk is used to reconstruct information in the event of a disk failure.

[0006] RAID level 4 systems are also comprised of one or more groups of N+1 disks wherein N disks are used to store data, and the additional disk is utilized to store parity information. RAID level 4 systems differ from RAID level 3 systems in that data to be saved is divided into larger portions, consisting of one or many blocks of data, for storage among the disks. Writes still require access to two disks, i.e., one of the N data disks and the

parity disk. In a similar fashion, read operations typically need only access a single one of the N data disks, unless the data to be read exceeds the block length stored on each disk. As with RAID level 3 systems, the parity disk is used to reconstruct information in the event of a disk failure.

[0007] RAID level 5 is similar to RAID level 4 except that parity information, in addition to the data, is distributed across the N+1 disks in each group. Although each group contains N+1 disks, each disk includes some blocks for storing data and some blocks for storing parity information. Where parity information is stored is controlled by an algorithm implemented by the user. As in RAID level 4 systems, RAID level 5 writes require access to at least two disks; however, no longer does every write to a group require access to the same dedicated parity disk, as in RAID level 4 systems. This feature provides the opportunity to perform concurrent write operations.

[0008] In a known disk array system the host operates as the RAID controller and performs the parity generation and checking. Having the host computer parity is expensive in host processing overhead. Furthermore, the known system includes a fixed data path structure interconnecting the plurality of disk drives with the host system. Rearrangement of the disk array system to accommodate different quantities of disk drives or different RAID configurations is not easily accomplished.

[0009] US-A-4,914,656 discloses a data storage system including a host bus means adapted to be connected to a host device, disk drive means adapted to be connected to a plurality of disk drive devices, selectively controllable coupling means connected to the host bus means and to a plurality of buses and which includes connecting means adapted to connect the host bus means to a first group of buses. EP-A-0,369,707 discloses a data storage system likewise including a host bus means for connection to a host device, disk drive means for connection to a plurality of disk drive devices and selectively controllable coupling means connected to the host bus means and the disk drive means.

[0010] EP-A-0,294,287 and EP-A-0,320,107 relate to data storage systems of the type defined within the present application but all of these aforementioned systems are disadvantageously restricted having regard to their adaptability and versatility for controlling data flow between a host and a disk array.

[0011] It is an object of the present invention to provide a data storage system of the kind specified which may be readily adapted to support various disk drive array configurations.

[0012] Therefore, according to the present invention, there is provided a data storage system including host bus means adapted to be connected to a host device, disk drive means adapted to be connected to a plurality of disk drive devices, selectively controllable coupling means connected to said host bus means and to a plurality of array buses coupled to said disk drive means,

first connecting means adapted to connect said host bus means to a first group of selected array buses characterized in that said first connecting means includes a plurality of register means each associated with a respective array bus and connected to said host bus means for receiving data therefrom, each register means having a respective bus driver connected thereto and also connected to the associated array bus, the bus drivers being selectively controllable to connect said host bus means to said first group of selected array buses, and in that said coupling means includes parity generation means, second connecting means adapted to connect a second group of selected array buses to the input of said parity generation means, and third connecting means adapted to connect the output of said parity generation means to a third group of selected array buses.

**[0013]** One embodiment of the invention will now be described by way of example, with reference to the accompanying drawings, in which:-

Figure 1 is a block diagram of a disk array controller incorporating the data and parity routing architecture of the present invention;

Figure 2 is a functional block diagram of the SCSI Array Data Path Chip (ADP) shown in Figure 1 which incorporates the data and parity routing architecture of the present invention;

Figure 3 is a block diagram of the DMA FIFO and routing blocks shown in Figure 2, representing a preferred embodiment of the present invention;

Figures 4A through 4F provide a more detailed block diagram of the architecture shown in Figure 3; and

Figures 5 through 18 illustrate some of the various configurations of the circuit of Figure 3. For example, Figure 8 shows the circuit configuration for performing a RAID level 3 write operation.

**[0014]** Referring now to Figure 1, there is seen a block diagram of a SCSI (small computer system interface) disk array controller incorporating the data and parity routing architecture of the present invention. The controller includes a host SCSI adapter 14 to interface between the host system data bus 12 and SCSI data bus 16. Parity generation and data and parity routing are performed within SCSI Array Data Path Chip (ADP) 10. ADP chip 10 interconnects SCSI data bus 16 with six additional data busses, identified by reference numerals 21 through 26. SCSI bus interface chips 31 through 35 interface between respective SCSI data busses 21 through 25 and corresponding external disk drive data busses 41 through 45. Bus 26 connects ADP chip 10 with a 64 kilobyte static random access memory (SRAM) 36. ADP chip 10, SCSI adapter 14 and SCSI bus interface chips 31 through 35 all operate under the control of a dedicated microprocessor 51.

**[0015]** All data busses shown, with the exception of bus 16, include nine data lines, one of which is a parity

bit line. Bus 16 includes 18 data lines, two of which are bus parity lines. Additional control, acknowledge and message lines, not shown, are also included with the data busses.

**[0016]** SCSI adapter 14, SCSI bus interface chips 31 through 36, SRAM 36 and microprocessor 51 are commercially available items. For example, SCSI adapter 14 may be a Fast SCSI 2 chip, SCSI bus interface chips 31 through 35 may be NCR 53C96 chips and microprocessor 51 could be a Motorola MC68020, 16 megahertz microprocessor. Also residing on the microprocessor bus are a one megabyte DRAM, a 128 kilobyte EPROM, an eight kilobyte EEPROM, a 68901 multifunction peripheral controller and various registers.

**[0017]** SCSI data path chip 10 is an application specific integrated circuit device capable of handling all data routing, data multiplexing and demultiplexing, and parity generation and checking aspects of RAID levels 1, 3 and 5. Data multiplexing of non-redundant data among five channels is also accommodated. ADP chip 10 handles the transfer of data between host SCSI adapter 14 and SCSI bus interface chips 31 through 35. The ADP chip also handles the movement of data to and from the 64 kilobyte SRAM during read/modify/write operations of RAID level 5.

**[0018]** Figure 2 is a functional block diagram of the SCSI Array Data Path Chip (ADP) shown in Figure 1. The ADP chip consists of the following internal functional blocks: control logic block 60 including diagnostic module 62, configuration control and status register module 64, interrupt module 66 and interface module 68; DMA FIFO block 70; and data and parity routing block 100.

**[0019]** The main function of control block 60 is to configure and to test the data path and parity path described in herein. Microprocessor interface module 68 contains basic microprocessor read and write cycle control logic providing communication and data transfer between microprocessor 51 (shown in Figure 1) and control and status registers within the control block. This interface utilizes a multiplexed address and data bus standard wherein microprocessor read and write cycles consist of an address phase followed by a data phase. During the address phase a specific register is selected by the microprocessor. During the following data phase, data is written to or read from the addressed register.

**[0020]** Configuration control and status register module 64 includes numerous eight-bit registers under the control of microprocessor interface module 68 as outlined above. The contents of the control registers determine the configuration of the data and parity paths. Status registers provide configuration and interrupt information to the microprocessor.

**[0021]** Diagnostic module 62 includes input and output data registers for the host and each array channel. The input registers are loaded by the processor and provide data to host bus 16, array busses 21 through 25 and SRAM bus 26 to emulate host or array transfers.

The output registers, which are loaded with data routed through the various data buses, are read by the microprocessor to verify the proper operation of selected data and parity path configurations.

**[0022]** Interrupt module 66 contains the control logic necessary to implement masking and grouping of ADP chip and disk controller channel interrupt signals. Any combination of five channel interrupt signals received from SCSI bus interface chips 31 through 35 and three internally generated ADP parity error interrupt signals can be combined together to generate interrupt signals for the microprocessor.

**[0023]** The function of DMA FIFO block 70 is to hold data received from the host until the array is ready to accept the data and to convert the data from eighteen bit bus 16 to nine bit busses 18. During read operations DMA FIFO block 70 holds the data received from the disk array until the host system is ready to accept the data and converts the data from nine bit busses 18 to eighteen bit bus 16.

**[0024]** Data and parity routing block 100 contains the steering logic for configuring the data and parity paths between the host, the disk array and SRAM 36 in response to control bytes placed into the control registers contained in module 64. Block 100 also includes logic for generating and checking parity, verifying data, broadcasting data, monitoring data transfers, and reconstructing data lost due to a single disk drive failure.

**[0025]** Figure 3 is a block diagram of DMA FIFO block 70 and routing block 100 shown in Figure 2, representing a preferred embodiment of the present invention. The data routing architecture includes five data channels, each channel providing a data path between SCSI adapter 14 and a corresponding one of SCSI bus interface chips 31 through 35. The first data channel includes a double register 101 connected between nine-bit host busses 16U and 16L and array bus 21. Busses 16U and 16L transfer sixteen bits of data and two bits of parity information between external SCSI adapter 14 and double register 101. Array bus 21 connects double register 101 with SCSI bus interface chip 31. Additional taps to array bus 21 will be discussed below. Data channels two through five are similarly constructed, connecting host busses 16U and 16L through respective double registers 102 through 105 and array busses 22 through 25 with SCSI bus interface chips 32 through 35.

**[0026]** Each one of double registers 101 through 105 includes an internal two-to-one multiplexer for selecting as input either one of host busses 16U or 16L when data is to be transferred from busses 16U and 16L to the array busses. Double registers 101 and 102 also each include an internal two-to-one multiplexer for selecting as input either the associated array bus or a parity bus 130 when data is to be transferred to host busses 16U and 16L.

**[0027]** An exclusive-OR circuit 110 is employed to generate parity information from data placed on array busses 21 through 25. A tap is provided between each

one of array busses 21 through 25 and a first input of a respective two-to-one multiplexer, the output of which forms an input to exclusive-OR circuit 110. The second inputs of each one of the multiplexers, identified by reference numerals 111 through 115, are connected to ground.

**[0028]** The output of exclusive-OR circuit 110 is provided via nine-bit bus 130 to double registers 101 and 102 and tristate buffers 121, 122, 123, 124, 125 and 126. Each one of tristate buffers 121 through 126, when enabled, provides the parity information generated by exclusive-OR circuit 110 to a respective one of busses 21 through 26. Bus 26 connects the output of tristate buffer 126 with SRAM 36 and one input of a two-to-one multiplexer 116. The output of multiplexer 116 is input to exclusive-OR circuit 110 to provide the architecture with parity checking and data reconstruction capabilities, to be discussed in greater detail below.

**[0029]** Also shown in Figure 3 are six transceivers, identified by reference numerals 141 through 146, for connecting respective busses 21 through 26 with processor bus 53 for transferring diagnostic information between the array channels and the processor. An additional transceiver 150 connects host busses 16U and 16L with processor bus 53. Transceivers 141 through 146 and 150 allow diagnostic operations by the controller processor.

**[0030]** Figures 4A through 4F provide additional detail concerning the construction and operation of double registers 101 through 105, transceivers 141 through 146 and transceiver 150, and control of data to and from SCSI bus interface chips 31 through 35 and SRAM 36.

**[0031]** Double register 101 is seen to include a multiplexer 203 for selecting data from either of busses 16U or 16L for placement into registers 205 and 207, a latch 209 and a buffer/driver 211 for placing the register contents onto array bus 21. Double register 101 further includes a multiplexer 213 for selecting data from either of busses 21 or 130 for placement into registers 215 and 217 and a buffer/driver 219 for placing the content of registers 213 and 215 onto bus 16L. Double register 102 is identical in construction to double register 101 for transferring data from either of busses 16U or 16L to bus 22, and from either of busses 21 or 130 to bus 16U. Double registers 103 and 104 are similar to double registers 101 and 102, respectively, except that neither includes a multiplexer corresponding to multiplexer 213 or a connection to bus 130. Double register 105 includes only those elements described above which are necessary to transfer information from busses 16U and 16L to bus 25.

**[0032]** Each one of transceivers 141 through 146 includes a register 223 connected to receive and store data residing on the associated array or SRAM bus, a buffer/driver 225 connected to provide the contents of register 223 to processor bus 53, a register 227 connected to receive and store data from processor bus 53, and a buffer/driver 229 connected to provide the contents of

register 227 to the associated array or SRAM bus. Transceiver 150 includes registers and drivers for storing and transferring data between processor bus 53 and each one of host busses 16U and 16L.

**[0033]** Array busses 21 through 25 each include a pair of parallel-connected buffer/drivers 233 and 235 for controlling the direction of data flow between the array bus and its associated SCSI bus interface chip. SRAM bus 26 is similarly connected with SRAM 36.

**[0034]** Control lines, not shown, are provided between control registers residing in configuration control and status module 64, discussed above, and double registers 101 through 105, multiplexers 111 through 116, tri-state buffers 121 through 126, and transceivers 141 through 146 and 150. Five data path control registers are responsible for configuration of the data path.

**[0035]** Data Path Control Register 1 controls data flow through double registers 101 through 105 during array write operations. Each control register bit controls the operation of one of registers 101 through 105 as identified below.

Data Path Control Register 1	
Bit 0	enables register 101
Bit 1	enables register 102
Bit 2	enables register 103
Bit 3	enables register 104
Bit 4	enables register 105
Bit 5	not used
Bit 6	not used
Bit 7	not used

**[0036]** When any one of control bits 0 through 4 are set high, data obtained from the host is transferred onto the internal array bus corresponding to the enabled register. The data obtained from the host is thereafter available to the SCSI bus interface chip connected to the array bus, the parity generating logic and to the diagnostic circuitry.

**[0037]** Data Path Control Register 2 controls the operation of multiplexers 111 through 116, thereby governing the input to exclusive-OR circuit 110. Each register bit controls the operation of one of multiplexers 111 through 116 as identified below.

Data Path Control Register 2	
Bit 0	selects input to MUX 111
Bit 1	selects input to MUX 112
Bit 2	selects input to MUX 113
Bit 3	selects input to MUX 114
Bit 4	selects input to MUX 115
Bit 5	selects input to MUX 116
Bit 6	not used
Bit 7	not used

**[0038]** In the table above, a logic one stored in one or more of register bits 0 through 5 sets the corresponding multiplexer to provide data from its associated array bus to exclusive-OR circuit 110. A logic zero stored in one or more of register bits 0 through 5 selects the corresponding multiplexer input connected to ground.

**[0039]** Data path control register 3 enables and disables the flow of data to or from SCSI bus interface chips 31 through 35 and SRAM 36. The functions of the individual register bits are described below.

Data Path Control Register 3	
Bit 0	enable SCSI DMA I/O on channel 1
Bit 1	enable SCSI DMA I/O on channel 2
Bit 2	enable SCSI DMA I/O on channel 3
Bit 3	enable SCSI DMA I/O on channel 4
Bit 4	enable SCSI DMA I/O on channel 5
Bit 5	enable data I/O with SRAM
Bit 6	not used
Bit 7	not used

**[0040]** A logic one stored in any one of bits 0 through 5 enables the corresponding channel. Data flow direction to or from SCSI bus interface chips 31 through 35 and SRAM 36 is determined by data path control register 5.

**[0041]** Data path control register 4, through associated chip hardware, enable specific RAID operations as described below.

Data Path Control Register 4	
Bit 0	array data transfer direction
Bit 1	RAID level 1 enable
Bit 2	RAID level 4 or 5 enable
Bit 3	RAID level 3 enable
Bit 4	drive configuration 2+1 enable
Bit 5	drive configuration 4+1 enable
Bit 6	array parity checking enable
Bit 7	not used

**[0042]** The direction of data flow through the register banks 101 through 105 is determined from the setting of bit 0. This bit is set high for host to array write operations and set low for array to host read operations. Bits 1, 2 and 3 enable the specific RAID operations when set high. Bits 4 and 5 enable the specific RAID level 3 operations when set high. Control bit 6 is set high to enable parity checking.

**[0043]** Data path control register 5 controls the data flow direction to or from SCSI bus interface chips 31 through 35 and SRAM 36.

Data Path Control Register 5	
Bit 0	array direction channel 1

(continued)

Data Path Control Register 5	
Bit 1	array direction channel 2
Bit 2	array direction channel 3
Bit 3	array direction channel 4
Bit 4	array direction channel 5
Bit 5	external SRAM direction
Bit 6	not used
Bit 7	not used

**[0044]** A logic one stored in any one of register bits 0 through 5 specifies that data is to be transferred from the array channel bus to its connected SCSI bus interface chip or SRAM. A logic 0 sets the data transfer direction from the SCSI bus interface chip or SRAM to the corresponding array channel bus. The control bits of register 3 must be set to enable SCSI bus interface chips 31 through 35 of SRAM 36 to perform a data transfer operation.

**[0045]** The five data path control registers discussed above enable or disable double registers 101 through 105, multiplexers 111 through 116, tri-state buffers 121 through 126, and transceivers 141 through 146 and 150 to configure the architecture to perform the various RAID levels 1, 3, 4 and 5 read and write operations which are described below with reference to Figures 5 through 14.

**[0046]** Figure 5 shows the data path architecture configured to perform RAID level 1 write operations on channels 1 and 2. The contents of data path control registers 1 through 5, expressed in hexadecimal notation, are: control register 1 - 03hex, control register 2 - 00hex, control register 3 - 03hex, control register 4 - 03hex, and control register 5 - 03hex. Double registers 101 and 102 are thereby enabled to provide data received from the host via busses 16U and 16L to busses 21 and 22. Data is transferred first from bus 16L to busses 21 and 22, and then from bus 16U to busses 21 and 22. SCSI bus interface chips 31 and 32 are enabled to transfer data from busses 21 and 22 to their associated array disks, each disk receiving the same data. Active busses (busses 16U, 16L, 21 and 22) are drawn with greater line weight in Figure 5. Arrows are provided next to the active busses to indicate the direction of data flow on the busses.

**[0047]** The RAID level 1 read configuration is shown in Figure 6. As both channel 1 and channel 2 array disks store the same information only one of the disks need be accessed for a RAID level 1 read. To read data from the channel 1 disk data path control register contents are set as follows: control register 1 - 00hex, control register 2 - 01hex, control register 3 - 01hex, control register 4 - 02hex, and control register 5 - 00hex. SCSI bus interface chip 31 is enabled to provide data from the channel 1 disk to bus 21. Multiplexer 111 is enabled to provide the data on bus 21 through parity generation circuit 110

to bus 130. Since multiplexers 112 through 116 are disabled the data provided to multiplexer 111 passes through the parity generator unaltered. Double registers 101 and 102 are enabled to provide data received from bus 130 to busses 16U and 16L, respectively. Bus 16U is provided with upper byte information while bus 16L transmits lower byte information.

**[0048]** Figure 7 shows the data path architecture configured to reconstruct the RAID level 1 channel 1 disk from the channel 2 disk. Data path control register contents are set as follows: control register 1 - 00hex, control register 2 - 02hex, control register 3 - 03hex, control register 4 - 03hex, and control register 5 - 01hex. SCSI bus interface chip 32 is enabled to provide data to bus 22, multiplexer 112 is enabled to provide bus 22 data to bus 130, tristate buffer 121 is enabled to place bus 130 data onto bus 21, and SCSI bus interface chip 31 is enabled to write data received from bus 21 onto the channel 1 disk. The architecture could similarly be configured to reconstruct the channel 2 disk from the channel 1 disk.

**[0049]** Figure 8 shows the data path architecture configured to permit RAID level 3, 4+1 (four data disks and one parity disk) write operations. The contents of the data path control registers are: control register 1 - 0Fhex, control register 2 - 0Fhex, control register 3 - 1Fhex, control register 4 - 29hex, and control register 5 - 1Fhex. Double registers 101 through 104 are enabled to provide data from the host to busses 21 through 24, respectively. Multiplexers 111 through 114 are enabled to provide data from busses 21 through 24 to parity generator 110, the output of which is provided via bus 130, enabled tri-state buffer 125, and bus 25 to bus interface 35. Bus interface chips 31 through 35 are enabled to provide data from busses 21 through 25 to their corresponding array disks.

**[0050]** The RAID level 3 read configuration is illustrated in Figure 9 where SCSI bus interface chips 31 through 34 are enabled to provide data from their corresponding array disks to double registers 101 through 104 via busses 21 through 25. Double registers 101 through 104 are enabled to provide the data received from busses 21 through 24 to busses 16U and 16L. The data path control registers are set as follows to configure the data paths: control register 1 - 00hex, control register 2 - 1Fhex, control register 3 - 1Fhex, control register 4 - 68hex, and control register 5 - 00hex. Parity checking is performed by exclusive-ORing the parity information stored on the channel 5 disk with the data read from the channel 1 through 4 disks. The output of parity generator 110 should be 00hex if the stored data is retrieved from the array without error.

**[0051]** Figure 10 shows the data path architecture configured to reconstruct the RAID level 3 channel 2 disk from the remaining data and parity disks. Data path control register contents are set as follows: control register 1 - 00hex, control register 2 - 1Dhex, control register 3 - 1Fhex, control register 4 - 29hex, and control reg-

ister 5 - 02hex. SCSI bus interface chips 31, 33, 34 and 35 and multiplexers 111, 113, 114 and 115 are enabled to provide data and parity information from the channel 1, 3, 4 and 5 array disks to parity generator 110. Parity generator 110 combines the received data and parity information to regenerate channel 2 data. The output of the parity generator is provided to the channel 2 disk through bus 130, enabled tristate device 122, bus 22 and enabled bus interface chip 32. Alternatively, the architecture could be configured to provide the reconstructed data directly to the host during read operations.

**[0052]** RAID level 5 write operations involve both a read and a write procedure. Figures 11 and 12 illustrate a RAID level 5 write involving the channel 1 and 2 array disks, wherein data is to be written to the channel 2 array disk and parity information is to be updated on the channel 1 disk. The data paths are first configured as shown in Figure 11 to read information from the channel 1 and 2 disks. This information is provided through multiplexers 111 and 112 to parity generator 110 and the result stored in external SRAM 36 via bus 130, enabled tristate buffer 136 and bus 26. The data path control registers contain the following codes: control register 1 - 00hex, control register 2 - 03hex, control register 3 - 23hex, control register 4 - 04hex, and control register 5 - 20hex.

**[0053]** New data and parity information is then written to the channel 1 and 2 array disks as shown in Figure 12. Double register 102 is enabled to receive the new data from the host and provide the data to the channel 2 disk via bus 22 and enabled SCSI bus interface chip 32. The new data is also provided to parity generator 110 through multiplexer 112. The information previously written into SRAM 36 is also provided to parity generator 110 through multiplexer 116. The output of the parity generator is the new parity which is provided through enabled tri-state buffer 121, bus 21 and enabled SCSI bus interface chip 31 to the channel 1 disk. The contents of the data path control registers for the second portion of the RAID level 5 write operation are as follows: control register 1 - 02hex, control register 2 - 22hex, control register 3 - 23hex, control register 4 - 05hex, and control register 5 - 03hex.

**[0054]** Figure 13 illustrates a RAID level 5 read operation. In this example the information to be read resides on the channel 5 array disk. The data path control registers are set to enable SCSI bus interface chip 35, multiplexer 115, and double registers 101 and 102 to transfer data from the channel 5 disk to the host. Control register contents are: control register 1 - 00hex, control register 2 - 10hex, control register 3 - 10hex, control register 4 - 04hex, and control register 5 - 00hex.

**[0055]** RAID level 5 disk reconstruction is similar to RAID level 3 disk reconstruction discussed above in connection with Figure 9. Figure 14 depicts a configuration for reconstructing data on channel 2. The data path control register contents are: control register 1 - 00hex, control register 2 - 1Dhex, control register 3 -

1Fhex, control register 4 - 25hex, and control register 5 - 02hex.

**[0056]** The data path architecture may also be configured to perform data verification, data broadcasting, and diagnostic operations. Additional control registers are provided to control transceivers 141 through 146 and transceiver 150 to configure the architecture to perform these additional operations. Figures 15 through 18 illustrate a few of the data verification, data broadcasting, and diagnostic operations which can be accomplished.

**[0057]** Figure 15 shows the data path architecture configured to verify that information stored on channel 1 and 2 array disks in accordance with RAID level 1 is correct. Channels 1 and 2 are enabled to receive data from SCSI bus interface chips 31 and 32. Multiplexers 111 and 112 are enabled to provide this data to the exclusive-OR circuit 110. The output of circuit 110, which should be 00hex if the channel 1 and 2 disks contain duplicate information, is provided through bus 130, enabled tri-state buffer 125, bus 25, enabled transceiver 145 and bus 53 to the processor for evaluation.

**[0058]** Figure 16 illustrates the data path architecture configured to verify that information stored in the array in accordance with RAID levels 3 or 5 is correct. Channels 1 through 5 are enabled to receive data from SCSI bus interface chips 31 through 35. Multiplexers 111 through 115 are enabled to provide this data to the exclusive-OR circuit 110. The output of circuit 110, which should be 00hex if the data and parity information are in agreement, is provided through bus 130, enabled tristate buffer 126, bus 26 and enabled transceiver 146 to the processor for evaluation.

**[0059]** Figures 17 and 18 show the data path architecture configured for data broadcasting wherein the same data is written to each of the disks in the array. Data broadcasting may be performed to initialize all the array disks with a known data pattern. In Figure 17 data is received by double register 101 from the processor via bus 53 and enabled transceiver 150. In Figure 18 data is provided to double register 101 by the host system. In both cases double register 101 is enabled to provide the received data to SCSI bus interface chips 31 via bus 21. Multiplexer 111 is enabled to provide bus 21 data to exclusive-OR circuit 110. Multiplexers 112 through 116 remain disabled so that the output of circuit 110 is equivalent to the data received from bus 21. Tristate devices 112 through 115 are enabled to provide the output of circuit 110 to SCSI bus interface chips 32 through 35.

**[0060]** It can thus be seen that there has been provided by the present invention a versatile data path and parity path architecture for controlling data flow between a host system and disk array. Those skilled in the art will appreciate that the invention is not limited to the details of the foregoing embodiments. For example, the architecture need not be limited to five channels. In addition, configurations other than those described above are possible. Processors, processor interfaces, and bus in-

terfaces other than the types shown in the Figures and discussed above can be employed. For example, the data path architecture can be constructed with ESDI, IPI or EISA devices rather than SCSI devices.

## Claims

1. A data storage system, including host bus means (16) adapted to be connected to a host device, disk drive means (31-35, 41-45) adapted to be connected to a plurality of disk drive devices, selectively controllable coupling means (10) connected to said host bus means (16) and to a plurality of array buses (21-25) coupled to said disk drive means (31-35, 41-45), first connecting means (205, 207, 211) adapted to connect said host bus means (16) to a first group of selected array buses (21-25) characterized in that said first connecting means includes a plurality of register means (205, 207) each associated with a respective array bus (21-25) and connected to said host bus means (16) for receiving data therefrom, each register means (205, 207) having a respective bus driver (211) connected thereto and also connected to the associated array bus (21-25), the bus drivers (211) being selectively controllable to connect said host bus means (16) to said first group of selected array buses (21-25), and in that said coupling means (10) includes parity generation means (110), second connecting means (111-115) adapted to connect a second group of selected array buses (21-25) to the input of said parity generation means (110), and third connecting means (121-125) adapted to connect the output of said parity generation means (110) to a third group of selected array buses (21-25).
2. A data storage system according to Claim 1, characterized in that said parity generation means includes an exclusive-OR circuit (110).
3. A data storage system according to Claim 2, characterized in that said coupling means (10) includes fourth connecting means (213-219) adapted to selectively connect the output of said exclusive-OR circuit (110) to said host bus means (16).
4. A data storage means according to Claim 2 or 3, characterized in that said second connecting means includes respective multiplexers (111-115) associated with said array buses (21-25), each said multiplexer (111-115) having a first input connected to its associated array bus (21-25), a second input connected to a reference voltage source, an output connected to an input of said exclusive-OR circuit (110) and a control signal input.
5. A data storage system according to Claim 2, 3 or 4,

characterized in that said third connecting means includes respective bus drivers (121-125) associated with said array buses (21-25), each bus driver (121-125) having an input connected to the output of said exclusive-OR circuit (110), an output connected to the associated array bus (21-25) and a control signal input.

6. A data storage system according to any one of Claims 2 to 5, characterized by temporary storage means (36), fifth connecting means (126) adapted to selectively connect the output of said exclusive-OR circuit (110) with said temporary storage means (36), and sixth connecting means (116) adapted to selectively connect said temporary storage means (36) to an input of said exclusive-OR circuit (110).
7. A data storage system according to Claim 6, characterized by a diagnostic bus (53), seventh connecting means (141-145) adapted to selectively connect said diagnostic bus (53) with a fourth group of selected array buses (21-25), and eighth connecting means (150) adapted to selectively connect said diagnostic bus (53) to said host bus means (16).
8. A data storage system according to any one of the preceding claims, characterized by a plurality of controllable interface devices (31035) adapted to connect said array buses (21-25) to respective disk drive buses (41-45).

## Patentansprüche

1. Datenspeichersystem, umfassend einen Host-Bus (16), der so ausgestaltet ist, daß er an ein Host-Gerät angeschlossen werden kann, Laufwerke (31-35, 41-45), die an eine Mehrzahl von Laufwerkgeräten angeschlossen werden können, selektiv steuerbare Kopplungsmittel (10), die mit dem genannten Host-Bus (16) und mit einer Mehrzahl von Array-Bussen (21, 25) verbunden sind, die mit den genannten Laufwerken (31-35, 41-45) verbunden sind, ein erstes Verbindungsmittel (205, 207, 211), das so ausgestaltet ist, daß es den genannten Host-Bus (16) mit einer ersten Gruppe von ausgewählten Array-Bussen (21-25) verbindet, dadurch gekennzeichnet, daß das genannte erste Verbindungsmittel eine Mehrzahl von Registern (205, 207) aufweist, die jeweils mit einem jeweiligen Array-Bus (21-25) assoziiert und mit dem genannten Host-Bus (16) verbunden sind, um Daten von diesem zu empfangen, wobei jedes Register (205, 207) einen jeweiligen Bustreiber (211) hat, der mit diesem sowie mit dem assoziierten Array-Bus (21-25) verbunden ist, wobei die Bustreiber (211) selektiv steuerbar sind, um den genannten Host-Bus (16) mit der ge-



nannten ersten Gruppe ausgewählter Array-Busse (21-25) zu verbinden, und dadurch, daß das genannte Kopplungsmittel (10) ein Paritätserzeugungsmittel (110), zweite Verbindungsmittel (111-115), die so ausgestaltet sind, daß sie eine zweite Gruppe von ausgewählten Array-Bussen (21-25) mit dem Eingang des genannten Paritätserzeugungsmittels (110) verbinden, und dritte Verbindungsmittel (121-125) umfaßt, die so ausgestaltet sind, daß sie den Ausgang des genannten Paritätserzeugungsmittels (110) mit einer dritten Gruppe von ausgewählten Array-Bussen (21-25) verbinden.

2. Datenspeichersystem nach Anspruch 1, dadurch gekennzeichnet, daß das genannte Paritätserzeugungsmittel eine Exklusiv-ODER-Schaltung (110) beinhaltet.
3. Datenspeichersystem nach Anspruch 2, dadurch gekennzeichnet, daß das genannte Kopplungsmittel (10) vierte Verbindungsmittel (213-219) aufweist, die so ausgestaltet sind, daß sie den Ausgang der genannten Exklusiv-ODER-Schaltung (110) mit dem genannten Host-Bus (16) verbinden.
4. Datenspeichersystem nach Anspruch 2 oder 3, dadurch gekennzeichnet, daß die genannten zweiten Verbindungsmittel jeweilige Multiplexer (111-115) aufweisen, die mit den genannten Array-Bussen (21-25) assoziiert sind, wobei jeder genannte Multiplexer (111-115) einen ersten Eingang, der mit seinem assoziierten Array-Bus (21-25) verbunden ist, einen zweiten Eingang, der mit einer Referenzspannungsquelle verbunden ist, einen Ausgang, der mit einem Eingang der genannten Exklusiv-ODER-Schaltung (110) verbunden ist, und einen Steuersignaleingang aufweist.
5. Datenspeichersystem nach Anspruch 2, 3 oder 4, dadurch gekennzeichnet, daß die genannten dritten Verbindungsmittel jeweilige Bustreiber (121-125) aufweisen, die mit den genannten Array-Bussen (21-25) assoziiert sind, wobei jeder Bustreiber (121-125) einen Eingang, der mit dem Ausgang der genannten Exklusiv-ODER-Schaltung (110) verbunden ist, einen Ausgang, der mit dem assoziierten Array-Bus (21-25) verbunden ist, und einen Steuersignaleingang aufweist.
6. Datenspeichersystem nach einem der Ansprüche 2 bis 5, gekennzeichnet durch ein temporäres Speichermittel (36), ein fünftes Verbindungsmittel (126), das so ausgestaltet ist, daß es den Ausgang der genannten Exklusiv-ODER-Schaltung (110) selektiv mit dem genannten temporären Speichermittel (36) verbindet, und ein sechstes Verbindungsmittel (116), das so ausgestaltet ist, daß es das genannte

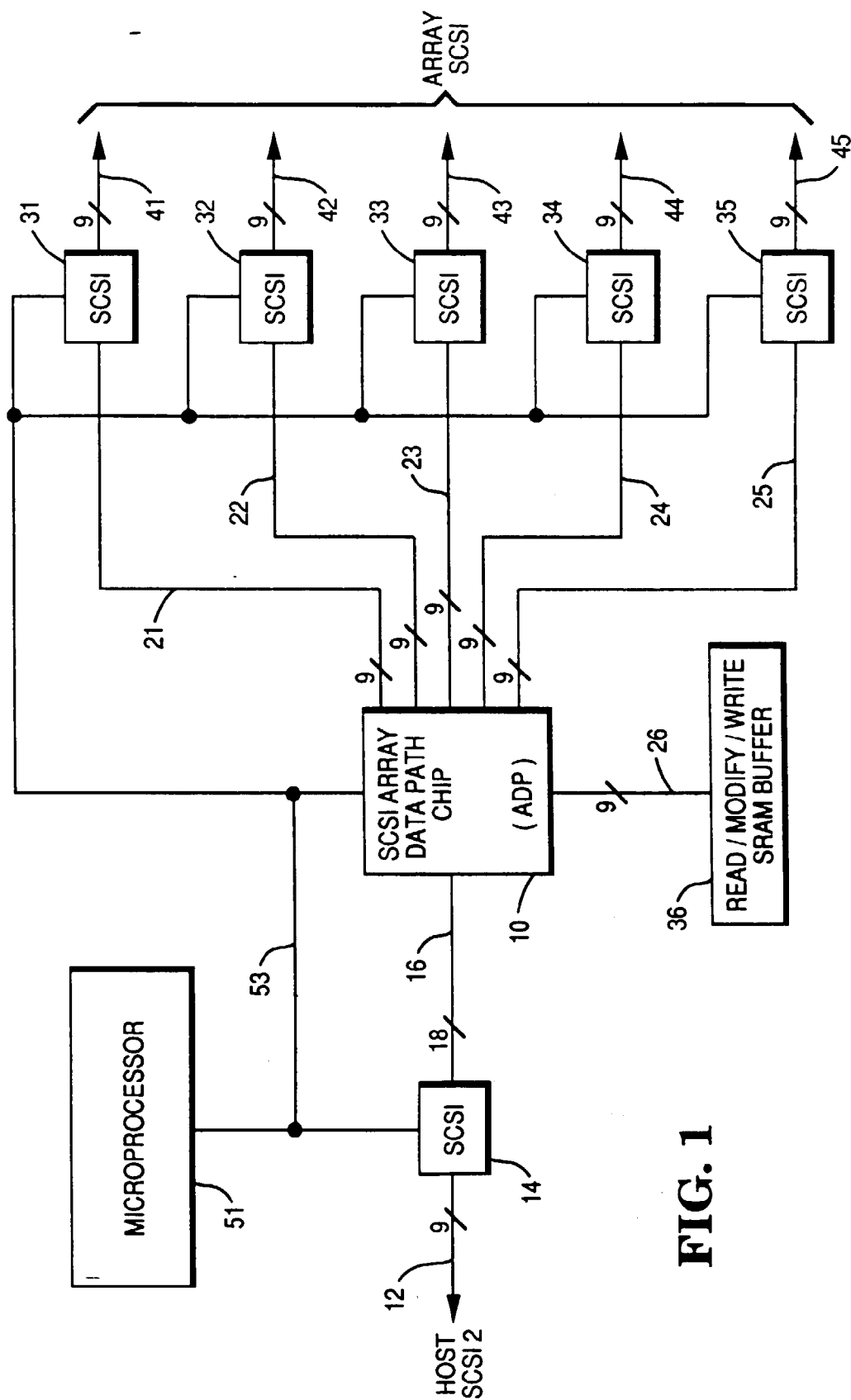
temporäre Speichermittel (36) selektiv mit einem Eingang der genannten Exklusiv-ODER-Schaltung (110) verbindet.

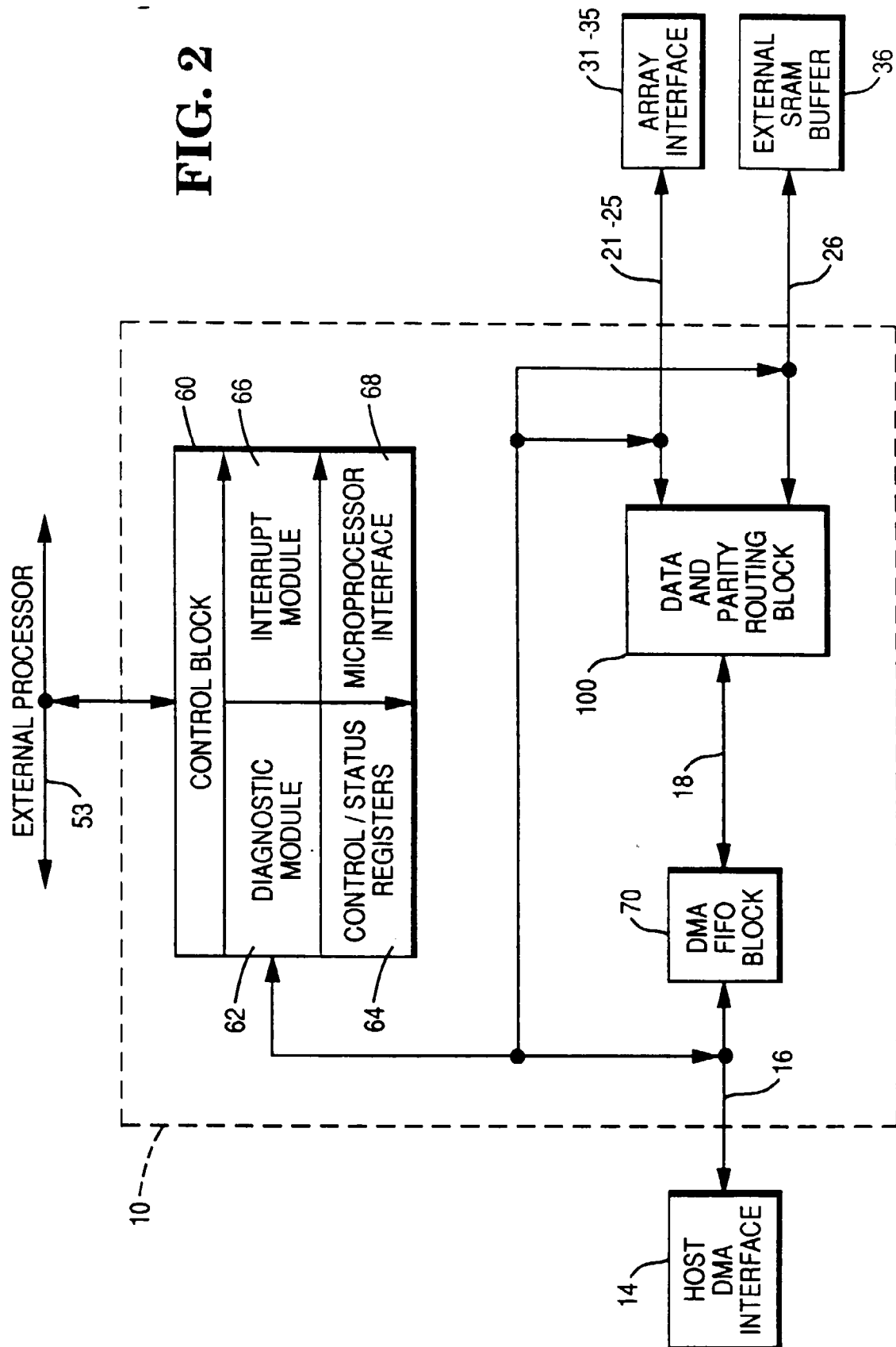
7. Datenspeichersystem nach Anspruch 6, gekennzeichnet durch einen Diagnosebus (53), siebte Verbindungsmittel (141-145), die so ausgestaltet sind, daß sie den genannten Diagnosebus (53) selektiv mit einer vierten Gruppe von ausgewählten Array-Bussen (21-25) verbinden, und ein achttes Verbindungsmittel (150), das so ausgestaltet ist, daß es den genannten Diagnosebus (53) selektiv mit dem genannten Host-Bus (16) verbindet.
8. Datenspeichersystem nach einem der vorherigen Ansprüche, gekennzeichnet durch eine Mehrzahl von steuerbaren Schnittstellengeräten (31035), die so ausgestaltet sind, daß sie die genannten Array-Busse (21-25) mit jeweiligen Laufwerksbussen (41-45) verbinden.

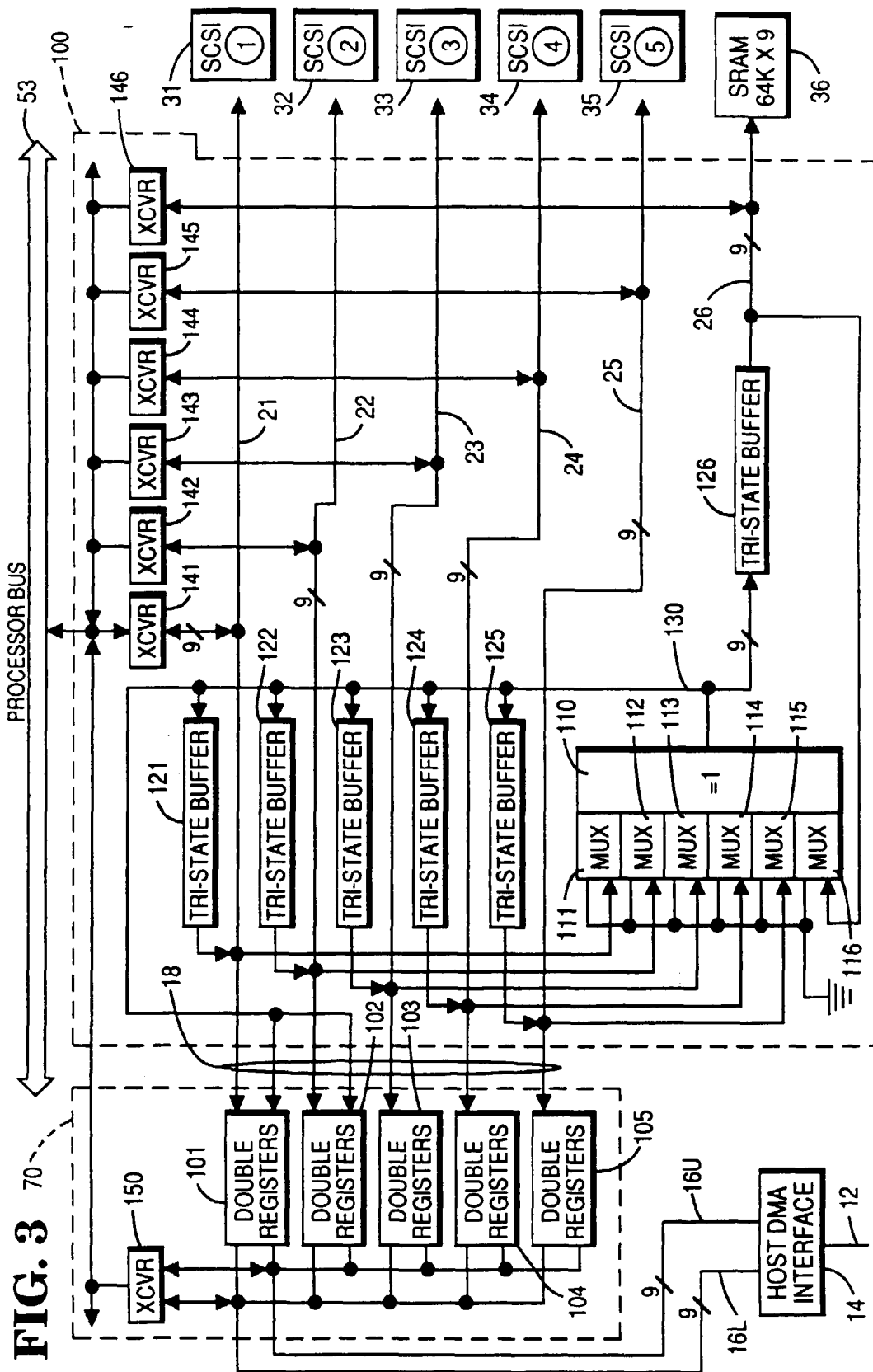
## Revendications

1. Système de stockage de données, incluant un moyen de bus hôte (16) adapté pour être connecté à un dispositif hôte, des moyens d'unités de disques (31-35, 41-45) adaptés pour être connectés à une pluralité de dispositifs d'unités de disques, un moyen de couplage sélectivement contrôlable (10) connecté audit moyen de bus hôte (16) et à une pluralité de bus de grappes (21-25) couplés auxdits moyens d'unités de disques (31-35, 41-45), un premier moyen de connexion (205, 207, 211) adapté pour connecter ledit moyen de bus hôte (16) à un premier groupe de bus de grappes sélectionnés (21-25), caractérisé en ce que ledit premier moyen de connexion inclut une pluralité de moyens de registre (205, 207) associés chacun à un bus de grappe respectif (21-25) et connectés audit moyen de bus hôte (16) pour recevoir des données de celui-ci, chaque moyen de registre (205, 207) ayant un pilote de bus respectif (211) connecté à ceux-ci et connecté également au bus de grappe associé (21-25), les pilotes de bus (211) étant sélectivement contrôlable pour connecter ledit moyen de bus hôte (16) audit premier groupe de bus de grappes sélectionnés (21-25), et en ce que ledit moyen de couplage (10) inclut un moyen de génération de parité (110), un deuxième moyen de connexion (111-115) adapté pour connecter un deuxième groupe de bus de grappes sélectionnés (21-25) à l'entrée dudit moyen de génération de parité (110), et un troisième moyen de connexion (121-125) adapté pour connecter la sortie dudit moyen de génération de parité (110) à un troisième groupe de bus de grappes sélectionnés (21-25).

2. Système de stockage de données selon la revendication 1, caractérisé en ce que ledit moyen de génération de parité inclut un circuit OU exclusif (110).  
respectifs (41-45).
3. Système de stockage de données selon la revendication 2, caractérisé en ce que ledit moyen de couplage (10) inclut un quatrième moyen de connexion (213-219) adapté pour connecter sélectivement la sortie dudit circuit OU exclusif (110) audit moyen de bus hôte (16).  
5  
10
4. Système de stockage de données selon la revendication 2 ou 3, caractérisé en ce que ledit deuxième moyen de connexion inclut des multiplexeurs respectifs (111-115) associés auxdits bus de grappes (21-25), chaque dit multiplexeur (111-115) ayant une première entrée connectée à son bus de grappe associé (21-25), une deuxième entrée connectée à une source de tension de référence, une sortie connectée à une entrée dudit circuit OU exclusif (110) et une entrée de signal de contrôle.  
15  
20
5. Système de stockage de données selon la revendication 2, 3 ou 4, caractérisé en ce que ledit troisième moyen de connexion inclut des pilotes de bus respectifs (121-125) associés auxdits bus de grappes (21-25), chaque pilote de bus (121-125) ayant une entrée connectée à la sortie dudit circuit OU exclusif (110), une sortie connectée au bus de grappes associé (21-25) et une entrée de signal de contrôle.  
25  
30
6. Système de stockage de données selon l'une quelconque des revendications 2 à 5, caractérisé par un moyen de stockage provisoire (36), un cinquième moyen de connexion (126) adapté pour connecter sélectivement la sortie dudit circuit OU exclusif (110) audit moyen de stockage provisoire (36), et un sixième moyen de connexion (116) adapté pour connecter sélectivement ledit moyen de stockage provisoire (36) à une entrée dudit circuit OU exclusif (110).  
35  
40
7. Système de stockage de données selon la revendication 6, caractérisé par un bus de diagnostic (53), un septième moyen de connexion (141-145) adapté pour connecter sélectivement ledit bus de diagnostic (53) à un quatrième groupe de bus de grappes sélectionnés (21-25), et un huitième moyen de connexion (150) adapté pour connecter sélectivement ledit bus de diagnostic (53) audit moyen de bus hôte (16).  
45  
50
8. Système de stockage de données selon l'une quelconque des revendications précédentes, caractérisé par une pluralité de dispositifs d'interface contrôlables (31035) adaptés pour connecter lesdits bus de grappes (21-25) à des bus d'unités de disques  
55







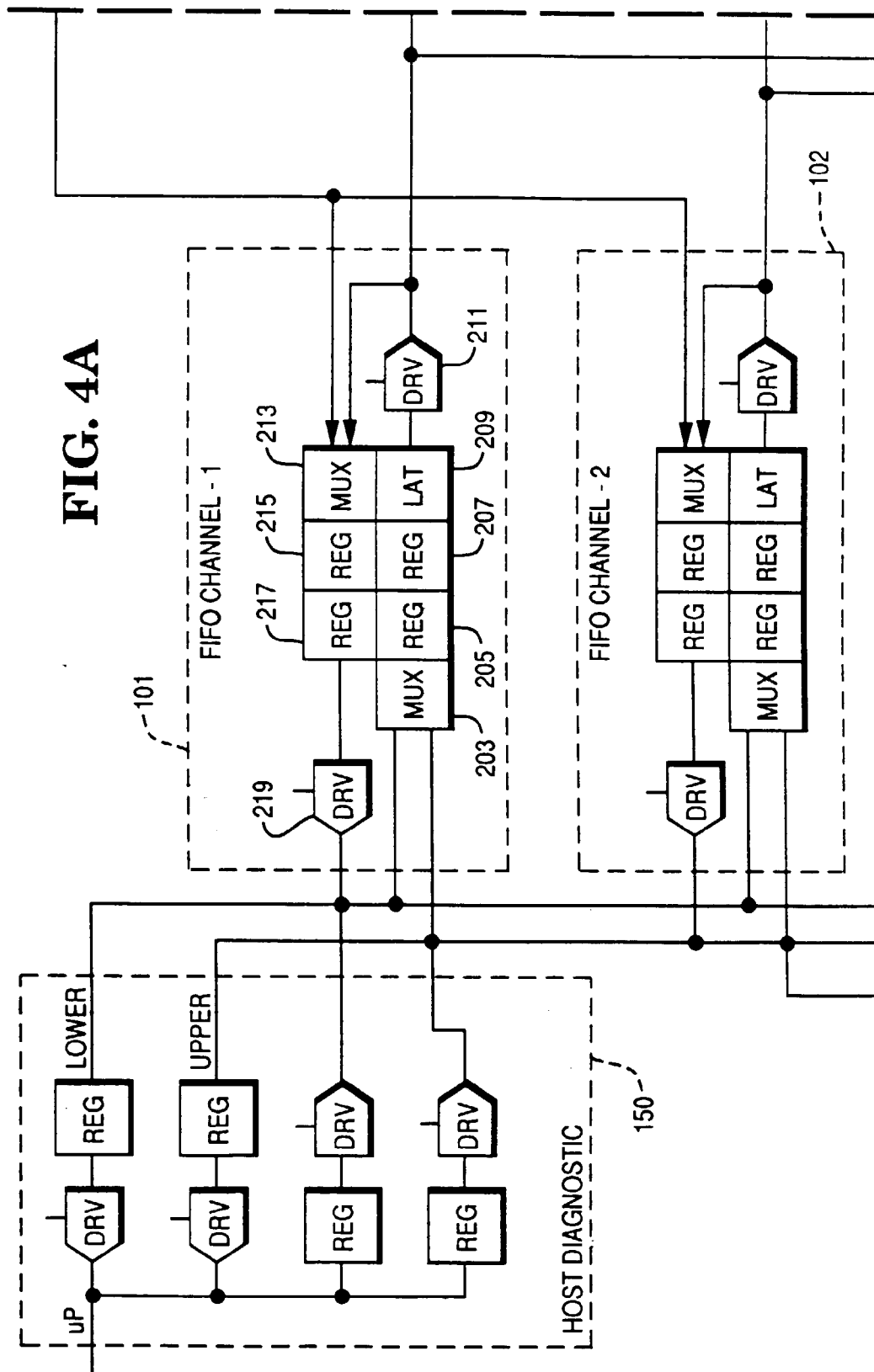
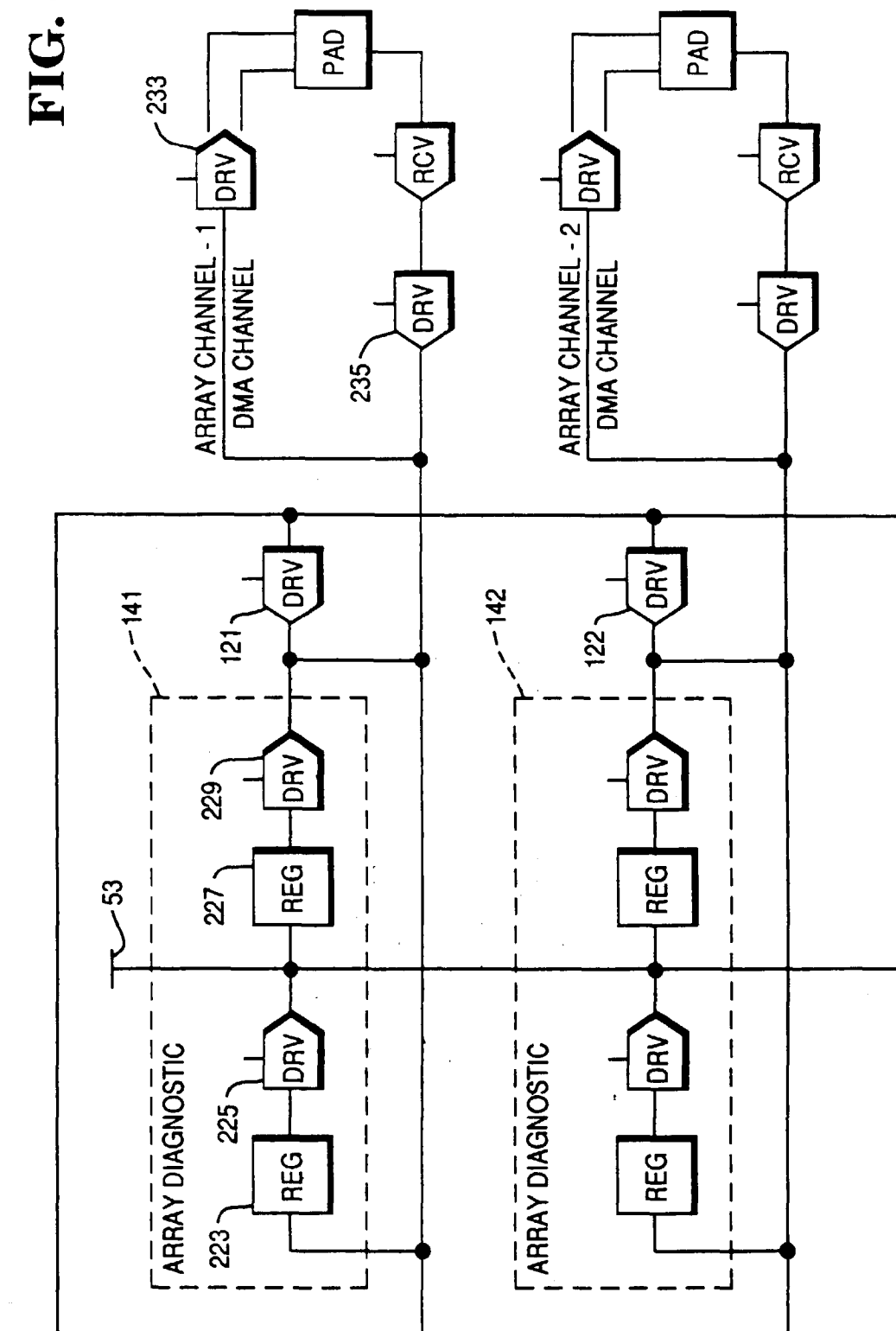
**FIG. 4A**

FIG. 4B



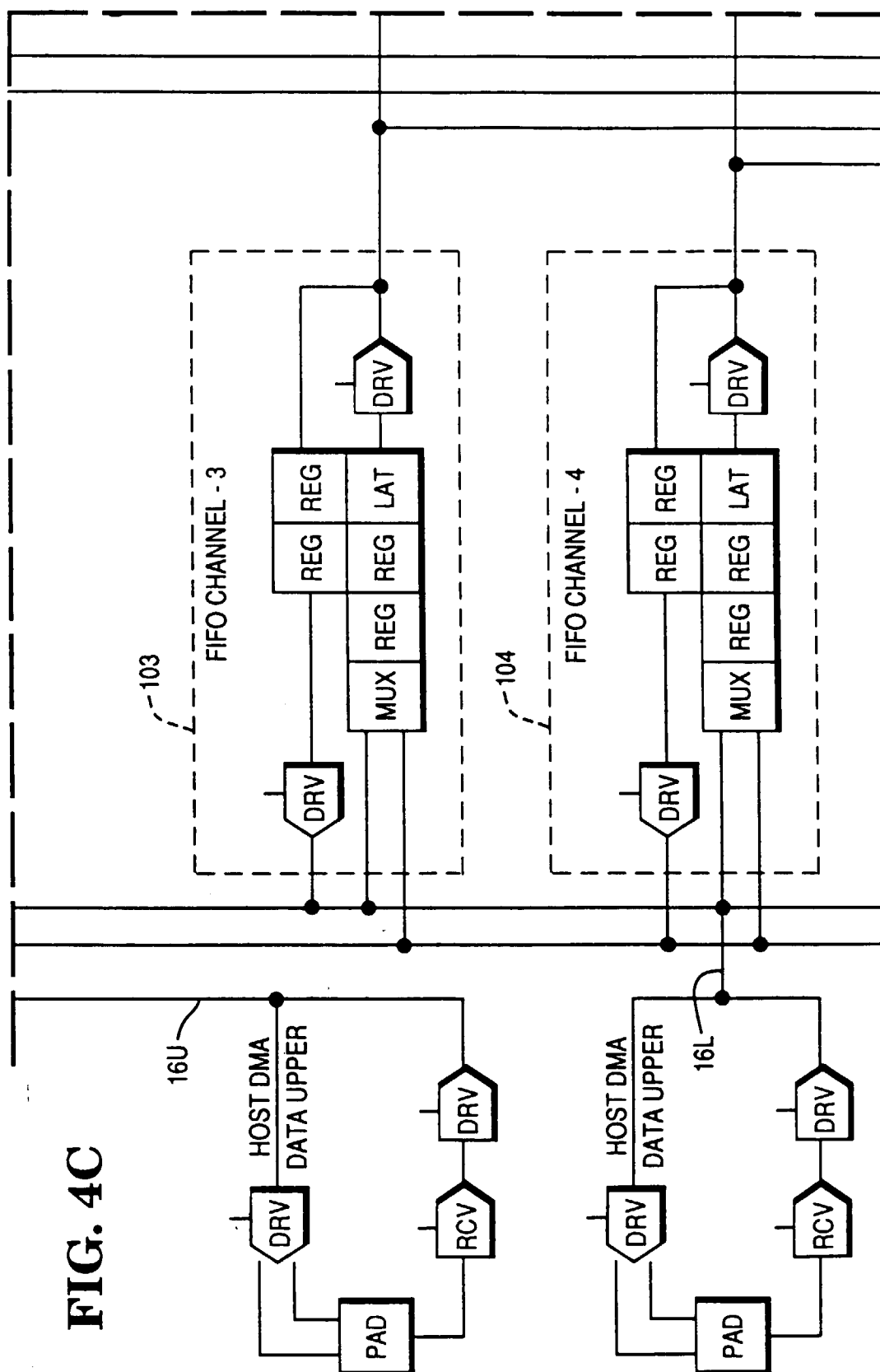
**FIG. 4C**



FIG. 4D

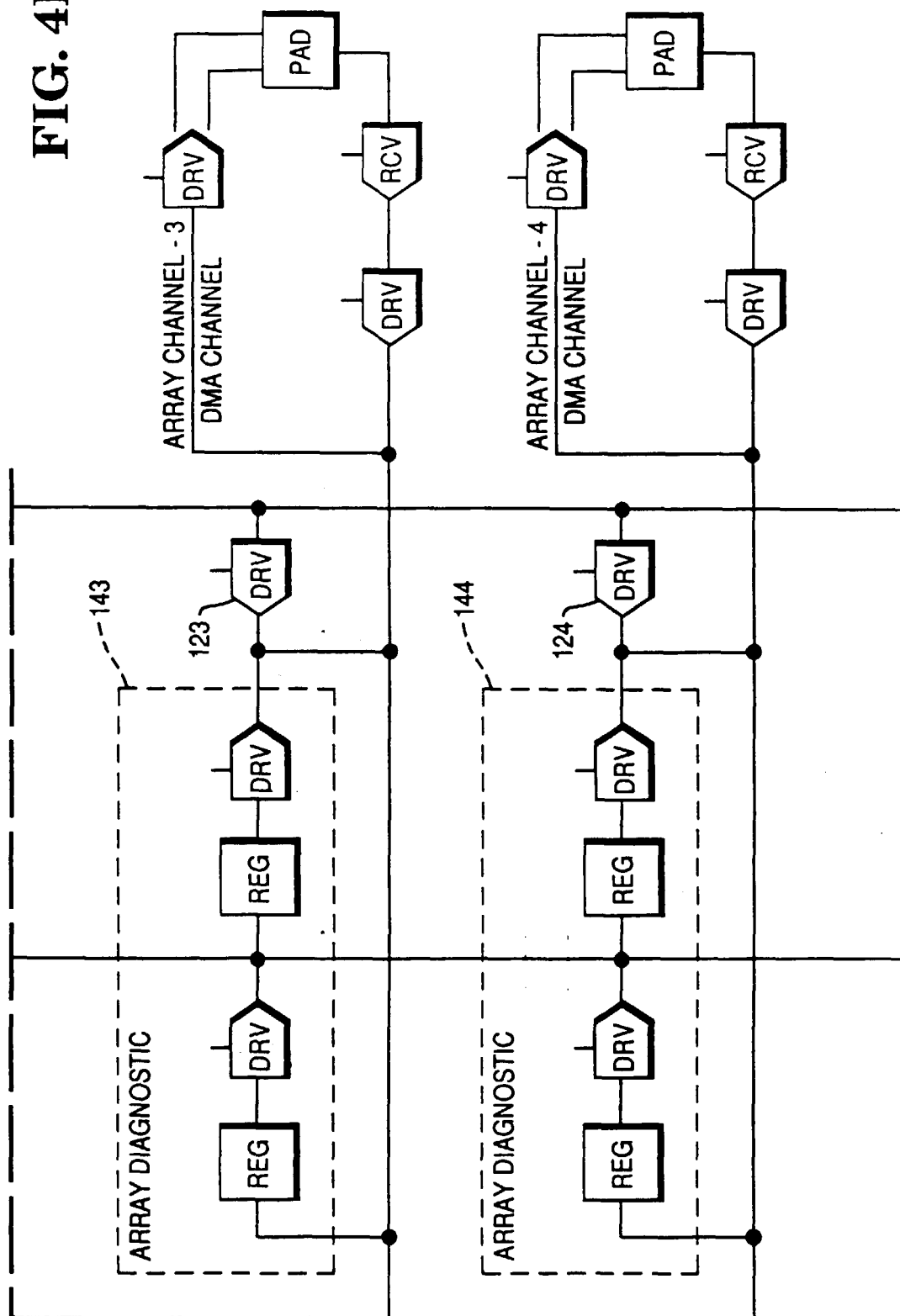


FIG. 4E

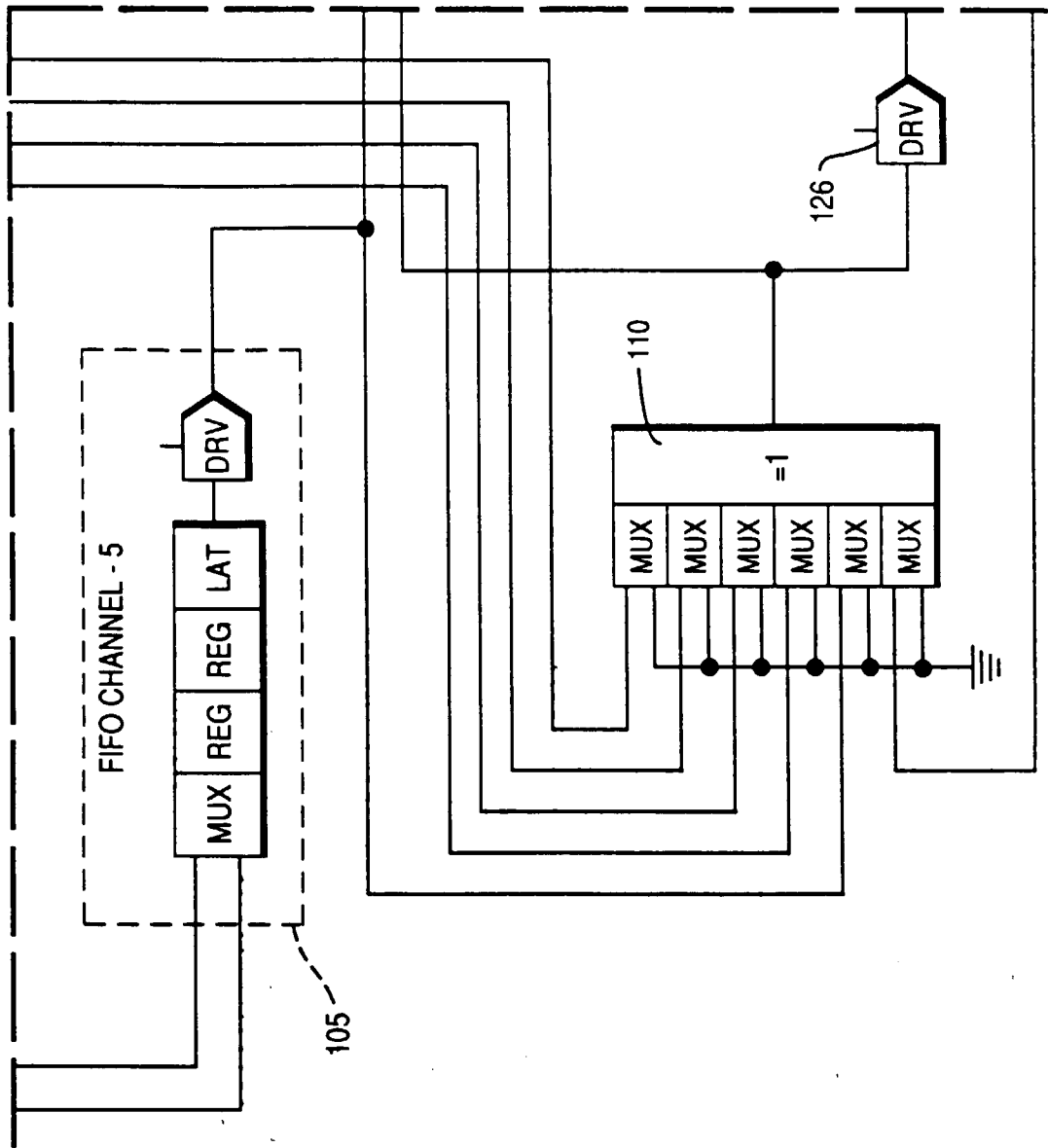


FIG. 4

FIG. 4A	FIG. 4B
FIG. 4C	FIG. 4D
FIG. 4E	FIG. 4F

**FIG. 4F**